

# Rigorous numerics for nonlinear operators with tridiagonal dominant linear part

Maxime Breden <sup>\*</sup>    Laurent Desvillettes <sup>†</sup>    Jean-Philippe Lessard <sup>‡</sup>

March 24, 2015

## Abstract

We present a method designed for computing solutions of infinite dimensional nonlinear operators  $f(x) = 0$  with a tridiagonal dominant linear part. We recast the operator equation into an equivalent Newton-like equation  $x = T(x) = x - Af(x)$ , where  $A$  is an approximate inverse of the derivative  $Df(\bar{x})$  at an approximate solution  $\bar{x}$ . We present rigorous computer-assisted calculations showing that  $T$  is a contraction near  $\bar{x}$ , thus yielding the existence of a solution. Since  $Df(\bar{x})$  does not have an asymptotically diagonal dominant structure, the computation of  $A$  is not straightforward. This paper provides ideas for computing  $A$ , and proposes a new rigorous method for proving existence of solutions of nonlinear operators with tridiagonal dominant linear part.

## Keywords

Tridiagonal operator · Contraction mapping · Rigorous numerics · Fourier series

## Mathematics Subject Classification (2010)

47H10 · 97N20 · 42A10 · 65L10 · 34B08

## 1 Introduction

Tridiagonal operators naturally arise in the theory of orthogonal polynomials, ordinary differential equations (ODEs), continued fractions, numerical analysis of partial differential equations (PDEs), integrable systems, quantum mechanics and solid state physics. Some differential operators can be represented by infinite tridiagonal matrices acting in sequence spaces, as it is the case for instance for differentiation in frequency space of the Hermite functions. Other examples come from the study of ODEs like the Mathieu equation, the spheroidal wave equation, the Whittaker-Hill equation and the Lamé equation.

While many well-developed methods and efficient algorithms already exist in the literature for solving linear tridiagonal matrix equations and computing their inverses, our own method has a different flavour. We aim at developing a computational method in order to

---

<sup>\*</sup>CMLA, ENS Cachan & CNRS, 61 avenue du Président Wilson, 94230 Cachan, France. [mbreden@ens-cachan.fr](mailto:mbreden@ens-cachan.fr)

<sup>†</sup>CMLA, ENS Cachan & CNRS, 61 avenue du Président Wilson, 94230 Cachan, France. [desville@cmla.ens-cachan.fr](mailto:desville@cmla.ens-cachan.fr)

<sup>‡</sup>Département de Mathématiques et de Statistique, Université Laval, 1045 avenue de la Médecine, Québec, QC, G1V0A6, Canada. [jean-philippe.lessard@mat.ulaval.ca](mailto:jean-philippe.lessard@mat.ulaval.ca)

prove, in a mathematically rigorous and constructive sense, existence of solutions to infinite dimensional nonlinear equations of the form

$$f(x) = \mathcal{L}(x) + N(x) = 0, \quad (1)$$

where  $\mathcal{L}$  is a tridiagonal linear operator and  $N$  is a nonlinear operator. The domain of the operator  $f$  is the space of algebraically decaying sequences

$$\Omega^s \stackrel{\text{def}}{=} \left\{ x = (x_k)_{k \geq 0} : \|x\|_s \stackrel{\text{def}}{=} \sup_{k \geq 0} \{|x_k| \omega_k^s\} < \infty \right\}, \quad (2)$$

where

$$\omega_k^s \stackrel{\text{def}}{=} \begin{cases} 1, & k = 0, \\ k^s, & k \geq 1. \end{cases}$$

The assumptions on the linear and nonlinear parts of (1) are that  $\mathcal{L} : \Omega^s \rightarrow \Omega^{s-s_L}$  and  $N : \Omega^s \rightarrow \Omega^{s-s_N}$ , for some  $s_L > s_N$ . Intuitively, this means that the linear part *dominates* the nonlinear part. Since  $\Omega^{s_1} \subset \Omega^{s_2}$  for  $s_1 > s_2$ , one can see that  $f$  maps  $\Omega^s$  into  $\Omega^{s-s_L}$ .

General nonlinear operator equations of the form  $f(x) = 0$  defined on the Banach space  $\Omega^s$  arise in the study of bounded solutions of finite and infinite dimensional dynamical systems. For instance,  $x = (x_k)_{k \geq 0}$  may be the infinite sequence of Fourier coefficients of a periodic solution of an ODE, a periodic solution of a delay differential equation (DDE) or an equilibrium solution of a PDE with Dirichlet, periodic or Neumann boundary conditions. The unknown  $x$  may also be the infinite sequence of Chebyshev coefficients of a solution of a boundary value problem (BVP), the Hermite coefficients of a solution of an ODE defined on an unbounded domain, or the Taylor coefficients of the solution of a Cauchy problem. In the case when the differential equation is smooth, the decay rate of the coefficients of  $x$  will be algebraic or even exponential [1]. In the present paper, we chose to solve (1) in the weighed  $\ell^\infty$  Banach space  $\Omega^s$  which corresponds to  $C^k$  solutions. In order to exploit the analyticity of the solutions, we could follow the idea of [2] and solve (1) in weighed  $\ell^1$  Banach spaces. This choice of space is not considered in the present paper.

Recently, several attempts to solve  $f(x) = 0$  in  $\Omega^s$  have been successful. They belong to a field now called *rigorous numerics*. This field aims at constructing algorithms that provide approximate solutions to a given problem, together with precise bounds implying the existence of an exact solution in the mathematically rigorous sense. Equilibria of PDEs [3, 4, 5], periodic solutions of DDEs [6], fixed points of infinite dimensional maps [7] and periodic solutions of ODEs [8, 9] have been computed using such methods.

One popular idea in rigorous numerics is to recast the problem  $f(x) = 0$  as a problem of fixed point of a Newton-like equation of the form  $T(x) = x - Af(x)$ , where  $A$  is an approximate inverse of  $Df(\bar{x})$ , and  $\bar{x}$  is a numerical approximation obtained by computing a finite dimensional projection of  $f$ . In [3, 4, 6, 7, 9, 5], the nonlinear equations under study have asymptotically diagonal or block-diagonal dominant linear part, which helps a lot in the computation of approximate inverses. In contrast, the present work considers problems with tridiagonal dominant linear part. To the best of our knowledge, this is the first attempt to compute rigorously solutions of such problems. While our proposed approach is designed for a specific class of operators (see assumptions (4) and (5)), we believe that it can be seen as a first step toward rigorously solving more complicated nonlinear operators with tridiagonal dominant linear part.

The paper is organized as follows. In Section 2, we present a method enabling to compute (with the help of the computer) pseudo-inverses of tridiagonal operators of a certain class. In Section 3, we recast the problem  $f(x) = 0$  as a fixed point problem  $T(x) = x - Af(x)$ , where

$A$  is a pseudo-inverse, and we present the rigorous computational method to prove existence of fixed points of  $T$ . In Section 4, we present an application and finally, in Section 5, we conclude by presenting some interesting future directions.

## 2 Computing pseudo-inverses of tridiagonal operators

This Section is devoted to the construction of a pseudo-inverse of a linear operator with tridiagonal tail (see (6)). We begin this Section by specifying the assumptions that we make on the growth of the tridiagonal terms. Then we use an LU-decomposition to formally obtain a formula for the pseudo-inverse. Finally, we check that the (formally defined) pseudo-inverse has good mapping properties (see Proposition 2.3).

Given three sequences  $(\lambda_k)_{k \geq 0}$ ,  $(\mu_k)_{k \geq 0}$ ,  $(\beta_k)_{k \geq 0}$  and  $x \in \Omega^s$ , we define the tridiagonal linear operator (acting on  $x$ )  $\mathcal{L}(x) = (\mathcal{L}_k(x))_{k \geq 0}$  of (1) by

$$\mathcal{L}_k(x) = \lambda_k x_{k-1} + \mu_k x_k + \beta_k x_{k+1}, \quad k \geq 1, \quad (3)$$

and  $\mathcal{L}_0(x) = \mu_0 x_0 + \beta_0 x_1$ . Assume that there exist real numbers  $s_L > 0$ ,  $0 < C_1 \leq C_2$  and an integer  $k_0$  such that

$$\forall k \geq 0, \quad \left| \frac{\lambda_k}{\omega_k^{s_L}} \right|, \left| \frac{\mu_k}{\omega_k^{s_L}} \right|, \left| \frac{\beta_k}{\omega_k^{s_L}} \right| \leq C_2 \quad \text{and} \quad \forall k \geq k_0, \quad C_1 \leq \left| \frac{\mu_k}{\omega_k^{s_L}} \right|. \quad (4)$$

Assume further the existence of  $\delta \in \left(0, \frac{1}{2}\right)$  and  $k_0 \geq 0$  such that

$$\forall k \geq k_0, \quad \left| \frac{\lambda_k}{\mu_k} \right|, \left| \frac{\beta_k}{\mu_k} \right| \leq \delta. \quad (5)$$

Then, under assumptions (4) and (5),  $\mathcal{L}$  defined by (3) is a tridiagonal operator which maps  $\Omega^s$  into  $\Omega^{s-s_L}$ . Indeed, if  $x \in \Omega^s$ , then

$$\begin{aligned} \|\mathcal{L}(x)\|_{s-s_L} &= \sup_{k \geq 0} \{|\mathcal{L}_k(x)| \omega_k^{s-s_L}\} \\ &\leq C_2 \left( \sup_{k \geq 1} \{|x_{k-1}| \omega_k^s\} + \sup_{k \geq 0} \{|x_k| \omega_k^s\} + \sup_{k \geq 0} \{|x_{k+1}| \omega_k^s\} \right) < \infty. \end{aligned}$$

From now on, assume for the sake of simplicity that  $s_N = 0$ , that is the nonlinear part  $N$  of (1) maps  $\Omega^s$  into  $\Omega^s$ . Since  $\Omega^s$  is an algebra under discrete convolutions when  $s > 1$  (e.g. see [5, 10]), then any  $N$  which is a combination of such convolutions maps  $\Omega^s$  into  $\Omega^s$ . Assume that using a finite dimensional projection  $f^{(m)} : \mathbb{R}^m \rightarrow \mathbb{R}^m$  of (1), we computed a numerical approximation  $\bar{x}$  such that  $f^{(m)}(\bar{x}) \approx 0$ . We identify  $\bar{x} \in \mathbb{R}^m$  and  $\bar{x} = (\bar{x}, 0, 0, 0, \dots) \in \Omega^s$ . We then try to construct a ball

$$B_{\bar{x}}(r) = \bar{x} + B_0(r) = \bar{x} + \{x \in \Omega^s : \|x\|_s \leq r\} = \{x \in \Omega^s : \|x - \bar{x}\|_s \leq r\}$$

centered at  $\bar{x}$  and containing a unique solution of (1), by showing that a specific Newton-like operator  $T(x) = x - Af(x)$  is a contraction on  $B_{\bar{x}}(r)$ . This requires the construction of an approximate inverse  $A$  of  $Df(\bar{x}) = \mathcal{L}(\bar{x}) + DN(\bar{x})$ . In order to do so, the structures of  $\mathcal{L}(\bar{x})$  and  $DN(\bar{x})$  need to be understood. From (3) and (4),  $\mathcal{L}(\bar{x})$  is a tridiagonal operator with entries growing to infinity at the rate  $k^{s_L}$ . Moreover, since  $DN(\bar{x})$  maps  $\Omega^s$  into  $\Omega^s$ , it is a

bounded linear operator. As mentioned above, the expectation is that the coefficients of  $\bar{x}$  decay fast to zero. This implies that a reasonable approximation  $A^\dagger$  of  $Df(\bar{x})$  is given by

$$A^\dagger \stackrel{\text{def}}{=} \begin{pmatrix} D & & & 0 \\ & \beta_{m-1} & & \\ & \lambda_m & \mu_m & \beta_m \\ 0 & & \lambda_{m+1} & \mu_{m+1} & \beta_{m+1} \end{pmatrix}, \quad (6)$$

with  $D \stackrel{\text{def}}{=} Df^{(m)}(\bar{x})$  for  $m$  large enough. We wish to find the inverse of  $A^\dagger$  in terms of  $D$ ,  $(\beta_k)_{k \geq m-1}$ ,  $(\mu_k)_{k \geq m}$  and  $(\lambda_k)_{k \geq m}$ . We assume therefore that

$$A^\dagger x = y, \quad (7)$$

where  $x$  and  $y$  are the infinite vectors

$$x = \begin{pmatrix} x_0 \\ x_1 \\ \cdot \\ \cdot \end{pmatrix}, \quad y = \begin{pmatrix} y_0 \\ y_1 \\ \cdot \\ \cdot \end{pmatrix}.$$

The infinite part of (7) writes

$$\begin{pmatrix} \mu_m & \beta_m & 0 & 0 & \dots \\ \lambda_{m+1} & \mu_{m+1} & \beta_{m+1} & 0 & \dots \\ 0 & \lambda_{m+2} & \mu_{m+2} & \beta_{m+2} & \dots \\ \cdot & \cdot & \cdot & \cdot & \dots \end{pmatrix} \begin{pmatrix} x_m \\ x_{m+1} \\ \cdot \\ \cdot \end{pmatrix} = \begin{pmatrix} y_m - \lambda_m x_{m-1} \\ y_{m+1} \\ \cdot \\ \cdot \end{pmatrix}. \quad (8)$$

We introduce the notations of the book of P.G. Ciarlet (see Theorem 4.3-2 on page 142 in [11]):

$$a_2 = \lambda_{m+1}, \quad a_3 = \lambda_{m+2}, \dots, \quad b_1 = \mu_m, \quad b_2 = \mu_{m+1}, \dots, \quad c_1 = \beta_m, \quad c_2 = \beta_{m+1}, \dots,$$

and  $(\delta_n)_{n \in \mathbb{N}}$  defined by the induction formula

$$\delta_0 = 1, \quad \delta_1 = b_1, \quad \text{and} \quad \delta_n = b_n \delta_{n-1} - a_n c_{n-1} \delta_{n-2}, \quad \text{for } n \geq 2.$$

Note that only the  $\delta_n$  are really useful.

Let us define the tridiagonal operator  $T$  by

$$T \stackrel{\text{def}}{=} \begin{pmatrix} b_1 & c_1 & 0 & 0 & \dots \\ a_2 & b_2 & c_2 & 0 & \dots \\ 0 & a_3 & b_3 & c_3 & \dots \\ \cdot & \cdot & \cdot & \cdot & \dots \end{pmatrix}. \quad (9)$$

For any infinite vector  $x = (x_0, \dots, x_k, \dots)^T$ , we introduce the notation

$$x_F \stackrel{\text{def}}{=} (x_0, \dots, x_{m-1})^T \quad \text{and} \quad x_I \stackrel{\text{def}}{=} (x_m, \dots, x_{m+k}, \dots)^T.$$

Using the notation  $\mathbf{e}_1 = (1, 0, 0, 0, \dots)^T$ , the system (8) becomes

$$Tx_I = y_I - \lambda_m x_{m-1} \mathbf{e}_1.$$

From Theorem 4.3-2 in [11], we compute an  $LU$ -decomposition of the tridiagonal operator defined in (9) as  $T = L_I U_I$ , where

$$L_I \stackrel{\text{def}}{=} \begin{pmatrix} 1 & 0 & 0 & \dots \\ a_2 \frac{\delta_0}{\delta_1} & 1 & 0 & \dots \\ 0 & a_3 \frac{\delta_1}{\delta_2} & 1 & \dots \\ \vdots & \vdots & \vdots & \dots \end{pmatrix} \quad \text{and} \quad U_I \stackrel{\text{def}}{=} \begin{pmatrix} \frac{\delta_1}{\delta_0} & c_1 & 0 & \dots \\ 0 & \frac{\delta_2}{\delta_1} & c_2 & \dots \\ 0 & 0 & \frac{\delta_3}{\delta_2} & \dots \\ \vdots & \vdots & \vdots & \dots \end{pmatrix}. \quad (10)$$

Hence, the system (8) becomes  $L_I z_I = y_I - \lambda_m x_{m-1} \mathbf{e}_1$  combined with  $U_I x_I = z_I$ , that is

$$\begin{pmatrix} 1 & 0 & 0 & \dots \\ a_2 \frac{\delta_0}{\delta_1} & 1 & 0 & \dots \\ 0 & a_3 \frac{\delta_1}{\delta_2} & 1 & \dots \\ \vdots & \vdots & \vdots & \dots \end{pmatrix} \begin{pmatrix} z_m \\ z_{m+1} \\ \vdots \\ \vdots \end{pmatrix} = \begin{pmatrix} y_m - \lambda_m x_{m-1} \\ y_{m+1} \\ \vdots \\ \vdots \end{pmatrix}, \quad (11)$$

combined with

$$\begin{pmatrix} \frac{\delta_1}{\delta_0} & c_1 & 0 & \dots \\ 0 & \frac{\delta_2}{\delta_1} & c_2 & \dots \\ 0 & 0 & \frac{\delta_3}{\delta_2} & \dots \\ \vdots & \vdots & \vdots & \dots \end{pmatrix} \begin{pmatrix} x_m \\ x_{m+1} \\ \vdots \\ \vdots \end{pmatrix} = \begin{pmatrix} z_m \\ z_{m+1} \\ \vdots \\ \vdots \end{pmatrix}. \quad (12)$$

Both infinite systems (11) and (12) can be explicitly solved.

System (11) leads to

$$z_m = y_m - \lambda_m x_{m-1},$$

and for any  $k \geq 1$

$$z_{m+k} = y_{m+k} + \sum_{l=1}^k (-1)^l a_{k-l+2} \dots a_{k+1} \frac{\delta_{k-l}}{\delta_k} y_{m+k-l} + (-1)^{k+1} a_2 \dots a_{k+1} \frac{\delta_0}{\delta_k} \lambda_m x_{m-1},$$

which we rewrite with infinite matrix/vectors notations as

$$z_I = L_I^{-1} [y_I - \lambda_m x_{m-1} \mathbf{e}_1] = L_I^{-1} y_I - \lambda_m x_{m-1} v_I, \quad (13)$$

where

$$z_I = \begin{pmatrix} z_m \\ z_{m+1} \\ z_{m+2} \\ \vdots \end{pmatrix}, \quad y_I = \begin{pmatrix} y_m \\ y_{m+1} \\ y_{m+2} \\ \vdots \end{pmatrix}, \quad v_I \stackrel{\text{def}}{=} L_I^{-1} \mathbf{e}_1 = \begin{pmatrix} 1 \\ -a_2 \frac{\delta_0}{\delta_1} \\ a_3 a_2 \frac{\delta_0}{\delta_2} \\ -a_4 a_3 a_2 \frac{\delta_0}{\delta_3} \\ \vdots \end{pmatrix}.$$

The second system (12) leads to the infinite sum (for any  $k \geq 0$ )

$$x_{m+k} = \frac{\delta_k}{\delta_{k+1}} z_{m+k} + \sum_{l=1}^{\infty} (-1)^l \frac{\delta_k}{\delta_{k+l+1}} c_{k+1} \dots c_{k+l} z_{m+k+l},$$

which we also rewrite with infinite matrix/vector notations as

$$x_I = U_I^{-1} z_I. \quad (14)$$

Coupling (13) and (14), we end up with

$$x_I = U_I^{-1} z_I = U_I^{-1} [L_I^{-1} y_I - \lambda_m x_{m-1} v_I] = U_I^{-1} L_I^{-1} y_I - \lambda_m x_{m-1} w_I, \quad (15)$$

where  $w_I \stackrel{\text{def}}{=} U_I^{-1} v_I$ . Denoting  $(U_I^{-1} L_I^{-1})_{r_0}$  the first row of the infinite matrix  $U_I^{-1} L_I^{-1}$  and  $(w_I)_0$  the first element of  $w_I$ , we can rewrite the first line of (15) as

$$x_m = (U_I^{-1} L_I^{-1})_{r_0} y_I - \lambda_m x_{m-1} (w_I)_0. \quad (16)$$

We now investigate the finite part of the linear system (7), which is given by

$$D \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{m-2} \\ x_{m-1} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \beta_{m-1} x_m \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{m-2} \\ y_{m-1} \end{pmatrix},$$

or, according to (16),

$$D \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{m-2} \\ x_{m-1} \end{pmatrix} + \beta_{m-1} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ (U_I^{-1} L_I^{-1})_{r_0} y_I - \lambda_m x_{m-1} (w_I)_0 \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{m-2} \\ y_{m-1} \end{pmatrix}.$$

Letting

$$K \stackrel{\text{def}}{=} D - \beta_{m-1} \lambda_m \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & (w_I)_0 \end{pmatrix},$$

we consider its inverse  $K^{-1}$ . We denote the last column of  $K^{-1}$  by  $(K^{-1})_{c_{m-1}}$ , its last row by  $(K^{-1})_{r_{m-1}}$ , and its last (“south-east”) element by  $(K^{-1})_{m-1, m-1}$ . Then we obtain

$$\begin{aligned} x_F &= K^{-1} y_F - \beta_{m-1} \left\{ (U_I^{-1} L_I^{-1})_{r_0} y_I \right\} (K^{-1})_{c_{m-1}} \\ &= K^{-1} y_F - \beta_{m-1} \left( \left\{ (K^{-1})_{c_{m-1}} \right\} \otimes \left\{ (U_I^{-1} L_I^{-1})_{r_0} \right\} \right) y_I, \end{aligned} \quad (17)$$

using the tensor product notation. The last line of this identity reads

$$x_{m-1} = (K^{-1})_{r_{m-1}} y_F - \beta_{m-1} \left\{ (U_I^{-1} L_I^{-1})_{r_0} y_I \right\} (K^{-1})_{m-1, m-1}. \quad (18)$$

Coming back to (15) and using (18), we see that

$$\begin{aligned}
x_I &= U_I^{-1} L_I^{-1} y_I - \lambda_m x_{m-1} w_I \\
&= U_I^{-1} L_I^{-1} y_I \\
&\quad - \lambda_m \left[ (K^{-1})_{r_{m-1}} y_F - \beta_{m-1} \left\{ (U_I^{-1} L_I^{-1})_{r_0} y_I \right\} (K^{-1})_{m-1, m-1} \right] w_I \\
&= U_I^{-1} L_I^{-1} y_I - \lambda_m w_I \left\{ (K^{-1})_{r_{m-1}} y_F \right\} \\
&\quad + \beta_{m-1} \lambda_m (K^{-1})_{m-1, m-1} w_I \left\{ (U_I^{-1} L_I^{-1})_{r_0} y_I \right\} \\
&= -\lambda_m \left( \left\{ w_I \right\} \otimes \left\{ (K^{-1})_{r_{m-1}} \right\} \right) y_F \\
&\quad + \left( U_I^{-1} L_I^{-1} + \beta_{m-1} \lambda_m (K^{-1})_{m-1, m-1} \left\{ w_I \right\} \otimes \left\{ (U_I^{-1} L_I^{-1})_{r_0} \right\} \right) y_I.
\end{aligned} \tag{19}$$

Putting together (17) and (19), we end up with

$$(A^\dagger)^{-1} = \begin{pmatrix} K^{-1} & -\beta_{m-1} \left( \left\{ (K^{-1})_{c_{m-1}} \right\} \otimes \left\{ (U_I^{-1} L_I^{-1})_{r_0} \right\} \right) \\ -\lambda_m \left\{ w_I \right\} \otimes \left\{ (K^{-1})_{r_{m-1}} \right\} & U_I^{-1} L_I^{-1} + \tilde{\Lambda} \end{pmatrix},$$

where

$$\tilde{\Lambda} \stackrel{\text{def}}{=} \beta_{m-1} \lambda_m (K^{-1})_{m-1, m-1} \left\{ w_I \right\} \otimes \left\{ (U_I^{-1} L_I^{-1})_{r_0} \right\}.$$

In order to get an approximate (pseudo) inverse of  $A^\dagger$ , we would like to get a numerical approximation of  $K^{-1}$ . However the definition of  $K$  involves  $(w_I)_0$ , which cannot be explicitly computed. By definition,  $w_I = U_I^{-1} L_I^{-1} \mathbf{e}_1$ , so using again the computations made in this Section, we get

$$\begin{aligned}
(w_I)_0 &= (U_I^{-1} v_I)_0 \\
&= \frac{\delta_0}{\delta_1} v_m + \sum_{l=1}^{\infty} (-1)^l \frac{\delta_0}{\delta_{l+1}} c_1 \dots c_l v_{m+l} \\
&= \frac{\delta_0}{\delta_1} + \sum_{l=1}^{\infty} \frac{\delta_0^2}{\delta_l \delta_{l+1}} c_1 \dots c_l a_2 \dots a_{l+1}.
\end{aligned}$$

Given a computational parameter  $L$ , we define

$$\tilde{w} \stackrel{\text{def}}{=} \frac{\delta_0}{\delta_1} + \sum_{l=1}^{L-1} \frac{\delta_0^2}{\delta_l \delta_{l+1}} c_1 \dots c_l a_2 \dots a_{l+1}, \tag{20}$$

and

$$\tilde{K} \stackrel{\text{def}}{=} D - \beta_{m-1} \lambda_m \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & \tilde{w} \end{pmatrix}.$$

We now can consider  $A_m$  a numerically computed inverse of  $\tilde{K}$  and then define the approximate (pseudo) inverse of  $A^\dagger$  as

$$A \stackrel{\text{def}}{=} \begin{pmatrix} A_m & -\beta_{m-1} \left( \left\{ (A_m)_{c_{m-1}} \right\} \otimes \left\{ (U_I^{-1} L_I^{-1})_{r_0} \right\} \right) \\ -\lambda_m \left\{ w_I \right\} \otimes \left\{ (A_m)_{r_{m-1}} \right\} & U_I^{-1} L_I^{-1} + \Lambda \end{pmatrix}, \quad (21)$$

where

$$\Lambda \stackrel{\text{def}}{=} \beta_{m-1} \lambda_m (A_m)_{m-1, m-1} \left\{ w_I \right\} \otimes \left\{ (U_I^{-1} L_I^{-1})_{r_0} \right\}.$$

**Lemma 2.1.** *Assume that  $m \geq k_0$  and  $\delta < \frac{1}{2}$ . Then  $U_I^{-1}$  maps  $\Omega^s$  into  $\Omega^{s+s_L}$ .*

*Proof.* Let  $z_I \in \Omega^s$  and  $x_I = U_I^{-1} z_I$ . Using (14) and the formula above, we get

$$\begin{aligned} |x_{m+k}| &\leq \frac{|\delta_k|}{|\delta_{k+1}|} |z_{m+k}| + \sum_{l=1}^{\infty} \frac{|\delta_k|}{|\delta_{k+l+1}|} |c_{k+1}| \dots |c_{k+l}| |z_{m+k+l}| \\ &\leq \frac{|\delta_k|}{|\delta_{k+1}|} |z_{m+k}| + \sum_{l=1}^{\infty} \delta^l \frac{|\delta_k|}{|\delta_{k+l+1}|} |b_{k+1}| \dots |b_{k+l}| |z_{m+k+l}|. \end{aligned} \quad (22)$$

Now remember that for all  $k \geq 2$ ,  $\delta_k = b_k \delta_{k-1} - a_k c_{k-1} \delta_{k-2}$ , so

$$\begin{aligned} \frac{|\delta_k|}{|\delta_{k-1}| |b_k|} &\geq 1 - \frac{|a_k| |c_{k-1}| |\delta_{k-2}|}{|b_k| |\delta_{k-1}|} \\ &\geq 1 - \frac{\delta^2 |b_{k-1}| |\delta_{k-2}|}{|\delta_{k-1}|}. \end{aligned}$$

We introduce  $u_k \stackrel{\text{def}}{=} \frac{|\delta_k|}{|\delta_{k-1}| |b_k|}$  which then satisfies

$$\begin{cases} u_1 = 1, \\ u_k \geq 1 - \frac{\delta^2}{u_{k-1}}, \quad \forall k \geq 2. \end{cases}$$

The study of the inductive sequence defined as above, but with  $\geq$  replaced by  $=$ , yields that for any  $k$ ,  $\gamma \leq u_k \leq 1$ , where  $\gamma \stackrel{\text{def}}{=} \frac{1}{2} + \sqrt{\frac{1}{4} - \delta^2}$  is the largest root of  $x = 1 - \frac{\delta^2}{x}$  (see Figure 1).



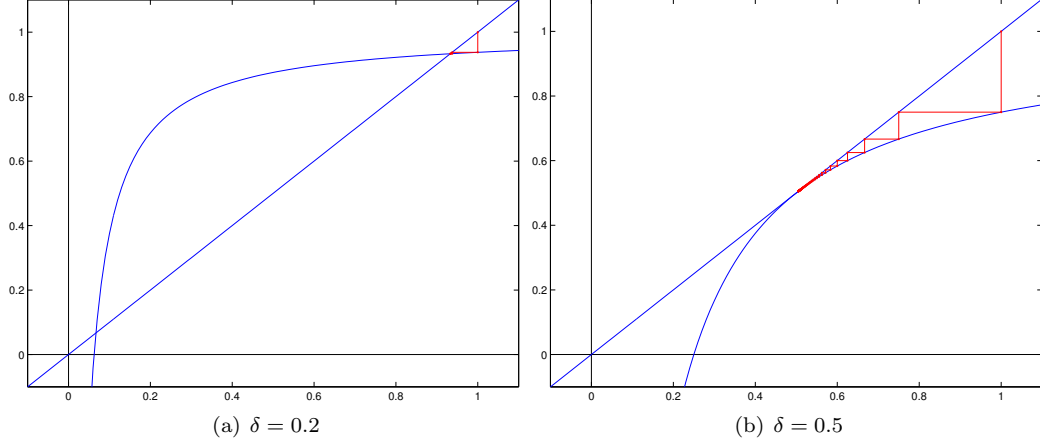


Figure 1: The iterations of  $u_{n+1} = 1 - \delta^2/u_n$  with  $u_1 = 1$ .

We can then rewrite (22) in order to get

$$\begin{aligned}
|x_{m+k}| &\leq \frac{|\delta_k|}{|\delta_{k+1}|} |z_{m+k}| + \sum_{l=1}^{\infty} \delta^l \frac{|\delta_k| \dots |\delta_{k+l}|}{|\delta_{k+1}| \dots |\delta_{k+l+1}|} |b_{k+1}| \dots |b_{k+l}| |z_{m+k+l}| \\
&\leq \frac{|\delta_k|}{|\delta_{k+1}|} |z_{m+k}| + \sum_{l=1}^{\infty} \delta^l \frac{1}{u_{k+1}} \dots \frac{1}{u_{k+l}} \frac{|\delta_{k+l}|}{|\delta_{k+l+1}|} |z_{m+k+l}| \\
&\leq \sum_{l=0}^{\infty} \left(\frac{\delta}{\gamma}\right)^l \frac{|\delta_{k+l}|}{|\delta_{k+l+1}|} |z_{m+k+l}| \\
&\leq \sum_{l=0}^{\infty} \left(\frac{\delta}{\gamma}\right)^l \frac{1}{\gamma |b_{k+l+1}|} |z_{m+k+l}| \\
&\leq \frac{\|z_I\|_s}{C_1 \gamma} \sum_{l=0}^{\infty} \left(\frac{\delta}{\gamma}\right)^l \frac{1}{(k+l+1)^{s_L} (m+k+l)^s}.
\end{aligned} \tag{23}$$

Finally, since  $\delta < \frac{1}{2} < \gamma$ ,

$$|x_{m+k}| (m+k)^{s+s_L} \leq \frac{\|z_I\|_s}{C_1 \gamma} \frac{1}{1 - \frac{\delta}{\gamma}} \frac{(m+k)^{s+s_L}}{(k+1)^{s_L} (m+k)^s}$$

and  $x_I \in \Omega^{s+s_L}$ . □

**Lemma 2.2.** Assume that  $m \geq k_0$  and  $\delta < \frac{1}{2}$ , Then  $L_I^{-1}$  maps  $\Omega^s$  into  $\Omega^s$ .

*Proof.* Let  $y_I \in \Omega^s$  and  $z_I = L_I^{-1} y_I$ . Using (13) and the formula above (without the last

term since we do not consider here  $L_I^{-1}(y_I - \lambda_m x_{m-1} \mathbf{e}_1)$ , we get

$$\begin{aligned}
|z_{m+k}| &\leq |y_{m+k}| + \sum_{l=1}^k \frac{|\delta_{k-l}|}{|\delta_k|} |a_{k-l+2}| \dots |a_{k+1}| |y_{m+k-l}| \\
&\leq |y_{m+k}| + \sum_{l=1}^k \delta^l \frac{|\delta_{k-l}|}{|\delta_k|} |b_{k-l+2}| \dots |b_{k+1}| |y_{m+k-l}| \\
&\leq |y_{m+k}| + \sum_{l=1}^k \delta^l \frac{|\delta_{k-l}| \dots |\delta_{k-1}|}{|\delta_{k-l+1}| \dots |\delta_k|} \frac{|b_{k-l+2}| |b_{k-l+3}| \dots |b_{k+1}|}{|b_{k-l+1}|} |y_{m+k-l}| \\
&\leq |y_{m+k}| + \sum_{l=1}^k \delta^l \frac{1}{u_{k-l+1}} \dots \frac{1}{u_k} \frac{|b_{k+1}|}{|b_{k-l+1}|} |y_{m+k-l}|,
\end{aligned}$$

where we use the sequence  $u_k$  introduced in the previous proof. We get

$$|z_{m+k}| \leq \sum_{l=0}^k \left(\frac{\delta}{\gamma}\right)^l \frac{|b_{k+1}|}{|b_{k-l+1}|} |y_{m+k-l}|, \quad (24)$$

and

$$\begin{aligned}
|z_{m+k}| (m+k)^s &\leq \frac{C_2 \|y\|_s}{C_1} \sum_{l=0}^k \left(\frac{\delta}{\gamma}\right)^l \left(\frac{k+1}{k+1-l}\right)^{s_L} \left(\frac{m+k}{m+k-l}\right)^s \\
&\leq \frac{C_2 \|y\|_s}{C_1} \sum_{l=0}^k \left(\frac{\delta}{\gamma}\right)^l \left(\frac{m+k}{k+1-l}\right)^{s+s_L}.
\end{aligned}$$

For any  $k \geq m$ , we then have

$$\begin{aligned}
|z_{m+k}| (m+k)^s &\leq \frac{2^{s+s_L} C_2 \|y\|_s}{C_1} \left( \sum_{l=0}^{\lfloor \frac{k}{2} \rfloor} \left(\frac{\delta}{\gamma}\right)^l \left(\frac{k}{k+1-l}\right)^{s+s_L} + \sum_{l=\lfloor \frac{k}{2} \rfloor+1}^k \left(\frac{\delta}{\gamma}\right)^l \left(\frac{k}{k+1-l}\right)^{s+s_L} \right) \\
&\leq \frac{2^{s+s_L} C_2 \|y\|_s}{C_1} \left( 2^{s+s_L} \sum_{l=0}^{\lfloor \frac{k}{2} \rfloor} \left(\frac{\delta}{\gamma}\right)^l + \left(\frac{\delta}{\gamma}\right)^{\frac{k}{2}} \sum_{l=\lfloor \frac{k}{2} \rfloor+1}^k k^{s+s_L} \right) \\
&\leq \frac{2^{s+s_L} C_2 \|y\|_s}{C_1} \left( \frac{2^{s+s_L}}{1 - \frac{\delta}{\gamma}} + \left(\frac{\delta}{\gamma}\right)^{\frac{k}{2}} \frac{k^{s+s_L+1}}{2} \right),
\end{aligned}$$

which is bounded uniformly in  $k$  since the last term goes to 0 when  $k$  goes to  $\infty$ , and the proof is complete.  $\square$

**Proposition 2.3.** Assume that  $m \geq k_0$  and  $\delta < \frac{1}{2}$ . Then  $A$  maps  $\Omega^s$  into  $\Omega^{s+s_L}$ .

*Proof.* Consider  $y = (y_F, y_I)^T \in \Omega^s$ . Let  $x = (x_F, x_I)^T = Ay$ . Then, by definition of the operator  $A$  in (21),

$$\begin{aligned}
x_F &= A_m y_F - \beta_{m-1} \left( \left\{ (A_m)_{c_{m-1}} \right\} \otimes \left\{ (U_I^{-1} L_I^{-1})_{r_0} \right\} \right) y_I \\
&= A_m y_F - \beta_{m-1} \left\{ (U_I^{-1} L_I^{-1})_{r_0} y_I \right\} (A_m)_{c_{m-1}}.
\end{aligned}$$

By the previous lemmas,  $U_I^{-1}L_I^{-1}y_I \in \Omega^{s+sL}$ , and in particular  $(U_I^{-1}L_I^{-1})_{r_0}y_I = (U_I^{-1}L_I^{-1}y_I)_0$  is well defined and so is  $x_F$ .

Using (21) again,

$$x_I = -\lambda_m \left( \left\{ w_I \right\} \otimes \left\{ (A_m)_{r_{m-1}} \right\} \right) y_F + U_I^{-1}L_I^{-1}y_I + \Lambda y_I.$$

Remember that  $w_I = U_I^{-1}L_I^{-1}\mathbf{e}_1$ , so that  $w_I \in \Omega^s$  for any  $s$ . According to the previous lemmas and the definition of  $\Lambda$  (see (21)), we see that  $x_I \in \Omega^{s+sL}$ .  $\square$

### 3 Computations of fixed points of the operator $T$

Our main motivation for computing approximate inverses is to prove existence, in a mathematically rigorous sense, of a fixed point of the Newton-like operator  $T$  in a set centered at a numerical approximation  $\bar{x}$ . The Newton-like operator has the form

$$T(x) = x - Af(x), \quad (25)$$

where  $A$  is the approximate inverse (21) of  $Df(\bar{x})$  computed using the theory of Section 2. Since  $f$  maps  $\Omega^s$  into  $\Omega^{s-sL}$  and  $A$  maps  $\Omega^s$  into  $\Omega^{s+sL}$  (thanks to Proposition 2.3), we see that  $T$  maps the Banach space  $\Omega^s$  into itself. Our goal is to obtain explicit bounds allowing us to show that a given  $T$  is a contraction on the ball  $B_{\bar{x}}(r)$ , which yields the existence of a fixed point of  $T$  (and thus of a zero of  $f$ ). The fixed point theorem that we use (see Theorem 3.1) requires bounds on  $T$  and its derivative. We get formulas for these bounds in Sections 3.2 and 3.3, and then explain in Section 3.4 how to use the so-called radii polynomials in order to find a radius  $r > 0$  such that  $T(B_{\bar{x}}(r)) \subset B_{\bar{x}}(r)$ , and such that  $T$  is a contraction on  $B_{\bar{x}}(r)$ .

Before proceeding further, we endow  $\Omega^s$  with the operation of discrete convolution. More precisely, given  $x = (x_k)_{k \geq 0}, y = (y_k)_{k \geq 0} \in \Omega^s$ , we extend  $x, y$  symmetrically by  $\tilde{x} = (x_k)_{k \in \mathbb{Z}}, \tilde{y} = (y_k)_{k \in \mathbb{Z}}$  where  $\tilde{x}_{-k} = x_k, \tilde{y}_{-k} = y_k$ , for  $k \geq 1$ . The discrete convolution of  $x$  and  $y$  is then denoted by  $x * y$ , and defined by the (infinite) sum

$$(x * y)_k = \sum_{\substack{k_1 + k_2 = k \\ k_1, k_2 \in \mathbb{Z}}} \tilde{x}_{k_1} \tilde{y}_{k_2}.$$

It is known that for  $s > 1$ ,  $(\Omega^s, *)$  is an algebra (e.g. see [10]), that is, if  $x, y \in \Omega^s$ , then  $x * y \in \Omega^s$ . This will be useful when we shall look for a bound such as (27) below. We start with a classical theorem, whose proof is standard (e.g. see the proof of Lemma 3.3 in [5]) and is a direct consequence of the contraction mapping theorem.

**Theorem 3.1.** *For a given  $s > 1$ , consider  $T: \Omega^s \rightarrow \Omega^s$  with  $T = (T_k)_{k \geq 0}$ ,  $T_k \in \mathbb{R}$ . Assume that there exists a point  $\bar{x} \in \Omega^s$  and vectors  $Y = \{Y_k\}_{k \geq 0}$  and  $Z = \{Z_k(r)\}_{k \geq 0}$ , with  $Y_k, Z_k(r) \in \mathbb{R}$ , satisfying (for all  $k \geq 0$ )*

$$|(T(\bar{x}) - \bar{x})_k| \leq Y_k, \quad (26)$$

and

$$\sup_{b_1, b_2 \in B_0(r)} \left| [DT(\bar{x} + b_1)b_2]_k \right| \leq Z_k(r). \quad (27)$$

*If there exists  $r > 0$  such that  $\|Y + Z(r)\|_s < r$ , then the operator  $T$  is a contraction in  $B_{\bar{x}}(r)$  and there exists a unique  $\hat{x} \in B_{\bar{x}}(r)$  such that  $T(\hat{x}) = \hat{x}$ .*

We shall see how to get the bounds  $Y$  (Section 3.2) and the bounds  $Z(r)$  (Section 3.3), and we shall provide an efficient way of finding a radius  $r > 0$  such that  $\|Y + Z(r)\|_s < r$  (Section 3.4). The first step however consists in looking for bounds on  $A$ . More precisely, we need some estimates in order to control the action of  $U_I^{-1}L_I^{-1}$ . This is the goal of the following Subsection.

### 3.1 Some preliminary computations

We introduce the notations

$$\theta \stackrel{\text{def}}{=} \frac{\delta}{\gamma} \quad \text{and} \quad \eta \stackrel{\text{def}}{=} \frac{1}{\gamma(1 - \theta^2)}. \quad (28)$$

**Lemma 3.2.** *Let  $y_I = (y_m, y_{m+1}, \dots)^T$  be an infinite vector and  $x_I = U_I^{-1}L_I^{-1}y_I$ . Assume that  $m \geq k_0$  and  $\delta < \frac{1}{2}$ . Then, for all  $k \geq 0$ ,*

$$|x_{m+k}| \leq \eta \left( \sum_{j=0}^k \theta^{k-j} \frac{|y_{m+j}|}{|\mu_{m+j}|} + \sum_{j=k+1}^{\infty} \theta^{j-k} \frac{|y_{m+j}|}{|\mu_{m+j}|} \right).$$

*Proof.* We again introduce  $z_I = L_I^{-1}y_I$ . Combining (23) from Lemma 2.1 and (24) from Lemma 2.2, we get

$$\begin{aligned} |x_{m+k}| &\leq \frac{1}{\gamma} \sum_{l=0}^{\infty} \sum_{j=0}^{k+l} \theta^{k+2l-j} \frac{|y_{m+j}|}{|b_{j+1}|} \\ &= \frac{1}{\gamma} \left( \sum_{j=0}^k \frac{|y_{m+j}|}{|b_{j+1}|} \sum_{l=0}^{\infty} \theta^{k+2l-j} + \sum_{j=k+1}^{\infty} \frac{|y_{m+j}|}{|b_{j+1}|} \sum_{l=j-k}^{\infty} \theta^{k+2l-j} \right) \\ &= \frac{1}{\gamma} \left( \sum_{j=0}^k \frac{|y_{m+j}|}{|b_{j+1}|} \frac{\theta^{k-j}}{1 - \theta^2} + \sum_{j=k+1}^{\infty} \frac{|y_{m+j}|}{|b_{j+1}|} \frac{\theta^{j-k}}{1 - \theta^2} \right) \\ &= \eta \left( \sum_{j=0}^k \theta^{k-j} \frac{|y_{m+j}|}{|\mu_{m+j}|} + \sum_{j=k+1}^{\infty} \theta^{j-k} \frac{|y_{m+j}|}{|\mu_{m+j}|} \right). \quad \square \end{aligned}$$

In particular, we immediately obtain the two following corollaries (always under the assumptions of Lemma 3.2) which will be useful in the sequel.

**Corollary 3.3.** *Recall (28). Then, for  $w_I = (w_m, w_{m+1}, \dots)^T \stackrel{\text{def}}{=} U_I^{-1}L_I^{-1}\mathbf{e}_1$ , we have*

$$|w_{m+k}| \leq \eta \theta^k \frac{1}{|\mu_m|}, \quad \text{for all } k \geq 0. \quad (29)$$

**Corollary 3.4.** *If  $y$  is such that  $y_{m+k} = 0$  for any  $k \geq n$ , then*

$$\forall k \leq n-2, \quad |x_{m+k}| \leq \eta \left( \sum_{l=0}^k \theta^{k-l} \frac{|y_{m+l}|}{|\mu_{m+l}|} + \sum_{l=k+1}^{n-1} \theta^{l-k} \frac{|y_{m+l}|}{|\mu_{m+l}|} \right) \quad (30)$$

and

$$\forall k \geq n-1, \quad |x_{m+k}| \leq \eta \theta^k \sum_{l=0}^{n-1} \frac{|y_{m+l}|}{\theta^l |\mu_{m+l}|}. \quad (31)$$

More generally, we will also need in the next two Subsections a uniform bound on  $|x_{m+k}|(m+k)^{s+s_L}$  for  $k$  large enough. We assume here that  $m \geq 2$  (which will always be the case in practice), and define for any integer  $M$

$$\chi = \chi(\theta, m, M, s, s_L) \stackrel{\text{def}}{=} \theta^{\frac{M}{2}} \frac{M}{2} \left( \frac{m+M}{m} \right)^{s+s_L} + \theta^{\sqrt{M}} \frac{M}{2} 2^{s+s_L} + \frac{1}{1-\theta} \left( \frac{m+M}{m+M-\sqrt{M}-1} \right)^{s+s_L}.$$

**Proposition 3.5.** *Suppose that  $M$  satisfies*

$$M \geq \max \left( \frac{-m \ln \sqrt{\theta} - s - s_L - 1 - \sqrt{(m \ln \sqrt{\theta} + s + s_L + 1)^2 - 4m \ln \sqrt{\theta}}}{2 \ln \sqrt{\theta}}, \frac{4}{(\ln \theta)^2}, m \right). \quad (32)$$

Then for all  $k < M$ ,

$$|x_{m+k}|(m+k)^{s+s_L} \leq \frac{\eta \|y_I\|_s}{C_1} \left( \sum_{l=0}^k \theta^{k-l} \left( \frac{m+k}{m+l} \right)^{s+s_L} + \frac{\theta}{1-\theta} \right), \quad (33)$$

and for all  $k \geq M$

$$|x_{m+k}|(m+k)^{s+s_L} \leq \frac{\eta \|y_I\|_s}{C_1} \left( \chi + \frac{\theta}{1-\theta} \right). \quad (34)$$

*Proof.* Thanks to Lemma 3.2,

$$\begin{aligned} |x_{m+k}|(m+k)^{s+s_L} &\leq \frac{\eta \|y_I\|_s}{C_1} \left( \sum_{l=0}^k \theta^{k-l} \left( \frac{m+k}{m+l} \right)^{s+s_L} + \sum_{l=k+1}^{\infty} \theta^{l-k} \left( \frac{m+k}{m+l} \right)^{s+s_L} \right) \\ &\leq \frac{\eta \|y_I\|_s}{C_1} \left( \sum_{l=0}^k \theta^{k-l} \left( \frac{m+k}{m+l} \right)^{s+s_L} + \frac{\theta}{1-\theta} \right). \end{aligned}$$

Then for  $k \geq M$ , we split the remaining sum

$$\begin{aligned} \sum_{l=0}^k \theta^{k-l} \left( \frac{m+k}{m+l} \right)^{s+s_L} &= \sum_{l=0}^{\lfloor \frac{k}{2} \rfloor - 1} \theta^{k-l} \left( \frac{m+k}{m+l} \right)^{s+s_L} + \sum_{l=\lfloor \frac{k}{2} \rfloor}^{k - \lfloor \sqrt{k} \rfloor - 1} \theta^{k-l} \left( \frac{m+k}{m+l} \right)^{s+s_L} + \sum_{l=k - \lfloor \sqrt{k} \rfloor}^k \theta^{k-l} \left( \frac{m+k}{m+l} \right)^{s+s_L} \\ &\leq \theta^{\frac{k}{2}} \frac{k}{2} \left( \frac{m+k}{m} \right)^{s+s_L} + \theta^{\sqrt{k}} \frac{k}{2} 2^{s+s_L} + \frac{1}{1-\theta} \left( \frac{m+k}{m+k-\sqrt{k}-1} \right)^{s+s_L} \\ &\leq \theta^{\frac{M}{2}} \frac{M}{2} \left( \frac{m+M}{m} \right)^{s+s_L} + \theta^{\sqrt{M}} \frac{M}{2} 2^{s+s_L} + \frac{1}{1-\theta} \left( \frac{m+M}{m+M-\sqrt{M}-1} \right)^{s+s_L} \\ &= \chi. \end{aligned}$$

The justification of the last inequality is contained in the following three lemmas.  $\square$

**Lemma 3.6.** *If  $M$  satisfies (32), then for all  $k \geq M$*

$$\theta^{\frac{k}{2}} \frac{k}{2} \left( \frac{m+k}{m} \right)^{s+s_L} \leq \theta^{\frac{M}{2}} \frac{M}{2} \left( \frac{m+M}{m} \right)^{s+s_L}.$$

*Proof.* For  $x > 0$ , let  $\varphi_1(x) \stackrel{\text{def}}{=} \theta^{\frac{x}{2}} x(m+x)^{s+s_L}$ , whose derivative is

$$\begin{aligned}\varphi_1'(x) &= \sqrt{\theta}^x \left( (\ln \sqrt{\theta}) x(m+x)^{s+s_L} + (m+x)^{s+s_L} + (s+s_L)x(m+x)^{s+s_L-1} \right) \\ &= (m+x)^{s+s_L-1} \sqrt{\theta}^x \left( (\ln \sqrt{\theta}) (m+x)x + (m+x) + (s+s_L)x \right) \\ &= (m+x)^{s+s_L-1} \sqrt{\theta}^x \left( (\ln \sqrt{\theta}) x^2 + (m \ln \sqrt{\theta} + s + s_L + 1)x + m \right).\end{aligned}$$

For  $0 < \theta < 1$ , the discriminant of  $\ln \sqrt{\theta} x^2 + (m \ln \sqrt{\theta} + s + s_L + 1)x + m$  given by

$$\Delta \stackrel{\text{def}}{=} \left( m \ln \sqrt{\theta} + s + s_L + 1 \right)^2 - 4m \ln \sqrt{\theta},$$

is positive. Since  $M$  satisfies (32),  $\varphi_1'(x) \leq 0$  for any  $x \geq M$  and so  $\varphi_1(k) \leq \varphi_1(M)$  for all  $k \geq M$ .  $\square$

**Lemma 3.7.** *If  $M$  satisfies (32), then for all  $k \geq M$ ,*

$$\theta^{\sqrt{k}} \frac{k}{2} 2^{s+s_L} \leq \theta^{\sqrt{M}} \frac{M}{2} 2^{s+s_L}.$$

*Proof.* Let  $\varphi_2(x) \stackrel{\text{def}}{=} \theta^{\sqrt{x}} x$ . Then

$$\varphi_2'(x) = \theta^{\sqrt{x}} \left( \frac{\ln \theta}{2\sqrt{x}} x + 1 \right) = \frac{\theta^{\sqrt{x}}}{2} (\sqrt{x} \ln \theta + 2).$$

Hence, for  $x \geq \frac{4}{(\ln \theta)^2}$ ,  $\varphi_2'(x) \leq 0$  and so  $\varphi_2(k) \leq \varphi_2(M)$  for all  $k \geq M$ .  $\square$

**Lemma 3.8.** *If  $M$  satisfies (32), then for all  $k \geq M$ ,*

$$\frac{1}{1-\theta} \left( \frac{m+k}{m+k-\sqrt{k}-1} \right)^{s+s_L} \leq \frac{1}{1-\theta} \left( \frac{m+M}{m+M-\sqrt{M}-1} \right)^{s+s_L}.$$

*Proof.* Let  $\varphi_3(x) \stackrel{\text{def}}{=} \frac{m+x}{m+x-\sqrt{x}-1}$ . Then

$$\varphi_3'(x) = \frac{m+x-\sqrt{x}-1-(m+x)\left(1-\frac{1}{2\sqrt{x}}\right)}{(m+x-\sqrt{x}-1)^2} = -\frac{x+2\sqrt{x}-m}{2\sqrt{x}(m+x-\sqrt{x}-1)^2}.$$

Hence, for  $x \geq m$ ,  $\varphi_3'(x) \leq 0$  and  $\varphi_3(k) \leq \varphi_3(M)$  for all  $k \geq M$ .  $\square$

Finally, we will need to bound the error made by using  $\tilde{w}$  instead of  $(w_I)_0$  for the definition (21) of  $A$ .

**Lemma 3.9.** *Assume that  $L \geq k_0$  and  $\delta < \frac{1}{2}$ . Then*

$$|(w_I)_0 - \tilde{w}| \leq \frac{\theta^{2L}}{|\mu_m|(1-\theta^2)}. \quad (35)$$

*Proof.* Using (5) together with the sequence  $(u_l)$  introduced in the proof of Lemma 2.1, we get

$$\begin{aligned}
|(w_I)_0 - \tilde{w}| &\leq \sum_{l=L}^{\infty} \frac{|\delta_0|^2}{|\delta_l| |\delta_{l+1}|} |c_1| \dots |c_l| |a_2| \dots |a_{l+1}| \\
&\leq \frac{|\delta_0|}{|\delta_1|} \sum_{l=L}^{\infty} \delta^{2l} \left( \frac{1}{u_1} \dots \frac{1}{u_l} \right) \left( \frac{1}{u_2} \dots \frac{1}{u_{l+1}} \right) \\
&\leq \frac{1}{|\mu_m|} \sum_{l=L}^{\infty} \theta^{2l} \\
&= \frac{\theta^{2L}}{|\mu_m| (1 - \theta^2)}. \quad \square
\end{aligned}$$

### 3.2 Computation of the $Y$ bounds

From now on, we shall assume for the sake of clarity that the nonlinearity  $N$  of  $f$  in (1) is a polynomial of degree two. The generalization to a polynomial nonlinearity of higher degree could be obtained thanks to the use of the estimates developed in [5] in order to bound terms like

$$(x^1 * \dots * x^p)_n$$

where  $x^1, \dots, x^p \in B_0(r)$ . Moreover, as long as one is interested in problems with nonlinearities built from elementary functions of mathematical physics (powers, exponential, trigonometric functions, rational, Bessel, elliptic integrals, etc.), our method is applicable. Indeed, since these nonlinearities are themselves solutions of low order linear or polynomial ODEs, they can be appended to the original problem of interest in order to obtain polynomial nonlinearities, albeit in a higher number of variables. This standard trick is explained in more details in [12], and is used in [18] to prove existence of periodic solutions in the planar circular restricted three body problem.

With this in mind, we are ready to compute the bound  $Y$  appearing in Theorem 3.1. In everything that follows,  $|\cdot|$ , when applied to vectors or matrices (even infinite dimensional), must be understood component-wise.

The main estimate of this subsection, that is the bound on  $Y$ , is presented in the following Proposition:

**Proposition 3.10.** *Consider an integer  $M$  such that*

$$M \geq \max \left( \frac{-s}{\ln \theta} - m, m - 2 \right), \quad (36)$$

*and define  $Y = (Y_k)_{k \geq 0}$  component-wise by*

$$Y_F \stackrel{\text{def}}{=} |A_m(f(\bar{x}))_F| + |\beta_{m-1}| \eta \left( \sum_{l=0}^{m-2} \theta^l \frac{|f(\bar{x})|_{m+l}}{|\mu_{m+l}|} \right) |(A_m)_{c_{m-1}}|, \quad (37)$$

$$\begin{aligned}
Y_{m+k} \stackrel{\text{def}}{=} & \left( |(A_m)_{r_{m-1}} f(\bar{x})_F| + |\beta_{m-1}| |(A_m)_{m-1, m-1}| \eta \left( \sum_{l=0}^{m-2} \theta^l \frac{|f(\bar{x})|_{m+l}}{|\mu_{m+l}|} \right) \right) \eta \theta^k \frac{|\lambda_m|}{|\mu_m|} \\
& + \eta \sum_{l=0}^k \theta^{k-l} \frac{|f(\bar{x})|_{m+l}}{|\mu_{m+l}|} + \eta \sum_{l=k+1}^{m-2} \theta^{l-k} \frac{|f(\bar{x})|_{m+l}}{|\mu_{m+l}|}, \quad \forall 0 \leq k \leq m-3, \quad (38)
\end{aligned}$$

$$\begin{aligned}
Y_{m+k} \stackrel{\text{def}}{=} & \left( \left| (A_m)_{r_{m-1}} f(\bar{x})_F \right| + \left| \beta_{m-1} (A_m)_{m-1, m-1} \right| \eta \left( \sum_{l=0}^{m-2} \theta^l \frac{|f(\bar{x})|_{m+l}}{|\mu_{m+l}|} \right) \right) \eta \theta^k \frac{|\lambda_m|}{|\mu_m|} \\
& + \eta \theta^k \sum_{l=0}^{m-2} \frac{|f(\bar{x})|_{m+l}}{\theta^l |\mu_{m+l}|}, \quad \forall m-2 \leq k \leq M,
\end{aligned} \tag{39}$$

and

$$Y_{m+k} \stackrel{\text{def}}{=} Y_{m+M} \frac{\omega_{m+M}^s}{\omega_{m+k}^s}, \quad \forall k > M. \tag{40}$$

Then

$$|T(\bar{x}) - \bar{x}| \leq Y.$$

*Proof.* By definition of  $T$ ,

$$|T(\bar{x}) - \bar{x}| = |Af(\bar{x})|.$$

Note that since we suppose that  $f$  is at most quadratic, and since  $\bar{x}$  is constructed in such a way that  $\bar{x}_k = 0$  for all  $k \geq m$ , we get the identity  $(f(\bar{x}))_{m+k} = 0$  for all  $k \geq m-1$ . Thanks to (21),

$$|(Af(\bar{x}))_F| \leq |A_m (f(\bar{x}))_F| + |\beta_{m-1}| |(U_I^{-1} L_I^{-1} (f(\bar{x}))_I)_0| |(A_m)_{c_{m-1}}|,$$

so that using (30) with  $n = m-1$  and  $k = 0$ , we get

$$|(Af(\bar{x}))_F| \leq |A_m (f(\bar{x}))_F| + |\beta_{m-1}| \eta \left( \sum_{l=0}^{m-2} \theta^l \frac{|f(\bar{x})|_{m+l}}{|\mu_{m+l}|} \right) |(A_m)_{c_{m-1}}|,$$

which provides the bound (37).

Using (21) again,

$$|(Af(\bar{x}))_I| \leq |\lambda_m| \left( \left| (A_m)_{r_{m-1}} f(\bar{x})_F \right| + \left| \beta_{m-1} (A_m)_{m-1, m-1} (U_I^{-1} L_I^{-1} f(\bar{x})_I)_0 \right| \right) |w_I| + |U_I^{-1} L_I^{-1} f(\bar{x})_I|,$$

so using (29), (30) and (31) (again with  $n = m-1$ ), we get

$$\begin{aligned}
|(Af(\bar{x}))_{m+k}| \leq & \left( \left| (A_m)_{r_{m-1}} f(\bar{x})_F \right| + \left| \beta_{m-1} (A_m)_{m-1, m-1} \right| \eta \left( \sum_{l=0}^{m-2} \theta^l \frac{|f(\bar{x})|_{m+l}}{|\mu_{m+l}|} \right) \right) \eta \theta^k \frac{|\lambda_m|}{|\mu_m|} \\
& + \eta \sum_{l=0}^k \theta^{k-l} \frac{|f(\bar{x})|_{m+l}}{|\mu_{m+l}|} + \eta \sum_{l=k+1}^{m-2} \theta^{l-k} \frac{|f(\bar{x})|_{m+l}}{|\mu_{m+l}|}, \quad \forall 0 \leq k \leq m-3,
\end{aligned}$$

which provides the bound (38), and

$$\begin{aligned}
|(Af(\bar{x}))_{m+k}| \leq & \left( \left| (A_m)_{r_{m-1}} f(\bar{x})_F \right| + \left| \beta_{m-1} (A_m)_{m-1, m-1} \right| \eta \left( \sum_{l=0}^{m-2} \theta^l \frac{|f(\bar{x})|_{m+l}}{|\mu_{m+l}|} \right) \right) \eta \theta^k \frac{|\lambda_m|}{|\mu_m|} \\
& + \eta \theta^k \sum_{l=0}^{m-2} \frac{|f(\bar{x})|_{m+l}}{\theta^l |\mu_{m+l}|}, \quad \forall k \geq m-2,
\end{aligned}$$

which provides the bound (39). Finally, by (36),  $\theta^k(m+k)^s \leq \theta^M(m+M)^s$  for all  $k > M$ , and we obtain the bound (40).  $\square$

We present in Section 3.4 the rationale behind the definition of  $Y_{m+k}$  for  $k > M$ .



### 3.3 Computation of the $Z$ bounds

In order to compute the  $Z$  bounds from Theorem 3.1, we need to estimate the quantity

$$DT(\bar{x} + y)z = (I - ADf(\bar{x} + y))z = (I - AA^\dagger)z - A(Df(\bar{x} + y) - A^\dagger)z$$

for all  $y, z \in B_0(r)$ . We are going to bound each term separately in the next two Sub-subsections. We introduce the notation

$$W_F^s \stackrel{\text{def}}{=} \left( \frac{1}{\omega_0^s}, \dots, \frac{1}{\omega_{m-1}^s} \right)^T. \quad (41)$$

#### 3.3.1 Estimates for $(I - AA^\dagger)z$

In this Sub-subsection, we present the bound on  $(I - AA^\dagger)z$ , which constitutes the first part of a bound for  $Z$ .

**Proposition 3.11.** *Let  $M$  be an integer satisfying (36). We define  $Z^1 = (Z_k^1)_{k \geq 0}$  component-wise by*

$$Z_F^1 \stackrel{\text{def}}{=} \left( \left| I - A_m \tilde{K} \right| W_F^s + \frac{|\beta_{m-1}| |\lambda_m| \theta^{2L}}{|\mu_m| \omega_{m-1}^s (1 - \theta^2)} |A_m|_{c_{m-1}} \right) r, \quad (42)$$

$$Z_{m+k}^1 \stackrel{\text{def}}{=} \left( \left| I - A_m \tilde{K} \right|_{r_{m-1}} W_F^s + \frac{|\beta_{m-1}| |\lambda_m| \theta^{2L}}{|\mu_m| \omega_{m-1}^s (1 - \theta^2)} |A_m|_{m-1, m-1} \right) \eta \theta^k \frac{|\lambda_m|}{|\mu_m|} r, \quad \forall 0 \leq k \leq M, \quad (43)$$

and

$$Z_{m+k}^1 \stackrel{\text{def}}{=} Z_{m+M}^1 \frac{\omega_{m+M}^s}{\omega_{m+k}^s}, \quad \forall k > M. \quad (44)$$

Then for all  $z \in B_0(r)$ ,

$$|(I - AA^\dagger)z| \leq Z^1.$$

*Proof.* Thanks to (6) and (21),

$$\begin{aligned} (AA^\dagger z)_F &= A_m \left( Dz_F + \begin{pmatrix} 0 \\ \vdots \\ \beta_{m-1} z_m \end{pmatrix} \right) - \beta_{m-1} (U_I^{-1} L_I^{-1} (Tz_I + \lambda_m z_{m-1} \mathbf{e}_1))_0 (A_m)_{c_{m-1}} \\ &= A_m Dz_F + \beta_{m-1} z_m (A_m)_{c_{m-1}} - \beta_{m-1} (z_m + \lambda_m z_{m-1} (w_I)_0) (A_m)_{c_{m-1}} \\ &= A_m \tilde{K} z_F + \beta_{m-1} \lambda_m (\tilde{w} - (w_I)_0) z_{m-1} (A_m)_{c_{m-1}}, \end{aligned}$$

and so

$$((I - AA^\dagger)z)_F = (I - A_m \tilde{K}) z_F + \beta_{m-1} \lambda_m (\tilde{w} - (w_I)_0) z_{m-1} (A_m)_{c_{m-1}}.$$

For  $z \in B_0(r)$  we have, using (35),

$$|(I - AA^\dagger)z|_F \leq \left( \left| I - A_m \tilde{K} \right| W_F^s + \frac{|\beta_{m-1}| |\lambda_m| \theta^{2L}}{|\mu_m| \omega_{m-1}^s (1 - \theta^2)} |A_m|_{c_{m-1}} \right) r,$$

which provides the bound (42).

Using again (6) and (21), we get

$$\begin{aligned}
(AA^\dagger z)_I &= -\lambda_m (A_m)_{r_{m-1}} \left( Dz_F + \begin{pmatrix} 0 \\ \vdots \\ \beta_{m-1} z_m \end{pmatrix} \right) w_I + (U_I^{-1} L_I^{-1} + \Lambda) (Tz_I + \lambda_m z_{m-1} \mathbf{e}_1) \\
&= z_I + \lambda_m w_I \\
&\quad \left( - (A_m)_{r_{m-1}} Dz_F - \beta_{m-1} (A_m)_{m-1, m-1} z_m + z_{m-1} + \beta_{m-1} (A_m)_{m-1, m-1} (z_I + \lambda_m z_{m-1} w_I)_0 \right) \\
&= z_I + \lambda_m \left( - (A_m)_{r_{m-1}} Dz_F + z_{m-1} + \beta_{m-1} \lambda_m (A_m)_{m-1, m-1} z_{m-1} (w_I)_0 \right) w_I \\
&= z_I + \lambda_m \left( z_{m-1} - (A_m)_{r_{m-1}} \tilde{K} z_F + \beta_{m-1} \lambda_m (A_m)_{m-1, m-1} z_{m-1} (\tilde{w} - (w_I)_0) \right) w_I \\
&= z_I + \lambda_m \left( \left( I - A_m \tilde{K} \right)_{r_{m-1}} z_F + \beta_{m-1} \lambda_m (A_m)_{m-1, m-1} z_{m-1} (\tilde{w} - (w_I)_0) \right) w_I,
\end{aligned}$$

and so

$$((I - AA^\dagger) z)_I = -\lambda_m \left( (I - A_m K)_{r_{m-1}} z_F + \beta_{m-1} \lambda_m (A_m)_{m-1, m-1} z_{m-1} (\tilde{w} - (w_I)_0) \right) w_I.$$

For  $z \in B_0(r)$  we have, using (29) and (35),

$$|(I - AA^\dagger) z|_{m+k} \leq \left( \left| I - A_m \tilde{K} \right|_{r_{m-1}} W_F^s + \frac{|\beta_{m-1}| |\lambda_m| \theta^{2L}}{|\mu_m| \omega_{m-1}^s (1 - \theta^2)} |A_m|_{m-1, m-1} \right) \eta \theta^k \frac{|\lambda_m|}{|\mu_m|} r, \quad \forall k \geq 0,$$

which gives (43), as well as (44) thanks to (36).  $\square$

### 3.3.2 Estimates for $A(Df(\bar{x} + y) - A^\dagger)z$

This Sub-subsection is devoted to the exposition of a bound for  $A(Df(\bar{x} + y) - A^\dagger)z$ , which constitutes the second (and last) part of a bound for  $Z$ . This bound is detailed in Proposition 3.18.

Recall the assumption that the nonlinear part  $N$  is polynomial of degree 2. Hence,  $Df(\bar{x} + y)$  can be written as a finite Taylor expansion

$$Df(\bar{x} + y) = Df(\bar{x}) + D^2f(\bar{x})(y),$$

and

$$(Df(\bar{x} + y) - A^\dagger)z = (Df(\bar{x}) - A^\dagger)z + D^2f(\bar{x})(y, z). \quad (45)$$

We are going to bound the two terms of (45) separately. Let us denote by  $\sigma$  the coefficient of degree 2 of  $f$ , that is  $D^2f(\bar{x})(y, z) = 2\sigma(y * z)$ . We bound this convolution product thanks to the following result:

**Lemma 3.12.** *Let  $s \geq 2$  be an algebraic decay rate and  $n \geq 6$ , let  $L \geq 1$  be computational parameters. For  $x, y \in \Omega^s$  and for any  $k \geq 0$ ,*

$$|(x * y)_k| \leq \alpha_k^s(n) \frac{\|x\|_s \|y\|_s}{\omega_k^s},$$

where

$$\alpha_k^s(n) \stackrel{\text{def}}{=} \begin{cases} 1 + 2 \sum_{l=1}^L \frac{1}{l^s} + \frac{2}{(s-1)L^{s-1}}, & k = 0, \\ 2 + 2 \sum_{l=1}^L \frac{1}{l^s} + \frac{2}{(s-1)L^{s-1}} + \sum_{l=1}^{k-1} \frac{k^s}{l^s(k-l)^s}, & 1 \leq k < n, \\ 2 + 2 \sum_{l=1}^L \frac{1}{l^s} + \frac{2}{(s-1)L^{s-1}} + 2 \left( \frac{n}{n-1} \right)^s + \left( \frac{4 \ln(n-2)}{n} + \frac{\pi^2 - 6}{3} \right) \left( \frac{2}{n} + \frac{1}{2} \right)^s, & k \geq n. \end{cases}$$

*Proof.* See [13] for a proof of this bound and [10] for a similar bound for  $1 < s < 2$ .  $\square$

**Remark 3.13.** It is important to notice here that  $\alpha_k^s(n) = \alpha_n^s(n)$  for all  $k \geq n$ . From now on, we assume that  $m$  is taken larger or equal to 6, which will allow us to use Lemma 3.12 with  $n = m$ . Note that this condition is not stringent, since in practice more than 6 modes are usually needed in order to get a good numerical solution  $\bar{x}$ .

We begin by bounding the first term of (45).

**Proposition 3.14.** Define  $C^1 = C^1(\bar{x}) = (C_k^1(\bar{x}))_{k \geq 0}$  component-wise by

$$C_0^1(\bar{x}) \stackrel{\text{def}}{=} 0, \quad C_k^1(\bar{x}) \stackrel{\text{def}}{=} 2|\sigma| \sum_{l=m-k}^{m-1} \frac{|\bar{x}_l|}{\omega_{k+l}^s}, \quad \forall 1 \leq k \leq m-1,$$

and

$$C_{m+k}^1(\bar{x}) \stackrel{\text{def}}{=} \frac{2|\sigma| \alpha_m^s(m) \|\bar{x}\|_s}{\omega_{m+k}^s}, \quad \forall k \geq 0.$$

Then for all  $z \in B_0(r)$

$$|(Df(\bar{x}) - A^\dagger)z| \leq C^1(\bar{x})r.$$

*Proof.* According to the definition of  $A^\dagger$  in (6), we see that

$$\begin{aligned} ((Df(\bar{x}) - A^\dagger)z)_F &= (Df(\bar{x})z)_F - Df^{(m)}(\bar{x})z_F - \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \beta_{m-1} z_m \end{pmatrix} \\ &= 2\sigma((\bar{x} * z)_F - (\bar{x} * z_F)_F), \end{aligned}$$

where in the convolution product,  $z_F$  must be understood as the infinite vector  $(z_F, 0, \dots, 0, \dots)^T$ . Therefore,  $((Df(\bar{x}) - A^\dagger)z)_0 = 0$ , and for all  $z \in B_0(r)$ ,

$$|(Df(\bar{x}) - A^\dagger)z|_k \leq 2|\sigma|r \sum_{l=m-k}^{m-1} \frac{|\bar{x}_l|}{\omega_{k+l}^s}, \quad \forall 1 \leq k \leq m-1.$$

Then, remembering that  $Df(\bar{x}) = \mathcal{L} + DN(\bar{x})$  and (6), we see that

$$((Df(\bar{x}) - A^\dagger)z)_I = (DN(\bar{x})z)_I = 2\sigma(\bar{x} * z)_I,$$

so that using Lemma 3.12, for all  $z \in B_0(r)$ , we end up with the bound

$$\left| (Df(\bar{x}) - A^\dagger) z \right|_{m+k} \leq \frac{2|\sigma| \alpha_{m+k}^s(m) \|\bar{x}\|_s}{\omega_{m+k}^s} r, \quad \forall k \geq 0. \quad \square$$

We now bound the second term of (45).

**Proposition 3.15.** *Recall (41) and define  $C^2 = (C_k^2)_{k \geq 0}$  component-wise by*

$$C_k^2 \stackrel{\text{def}}{=} \frac{2|\sigma| \alpha_k^s(m)}{\omega_k^s}, \quad \forall k \geq 0.$$

Then for all  $y, z \in B_0(r)$

$$\left| D^2 f(\bar{x})(y, z) \right| \leq C^2 r^2.$$

*Proof.* Remembering that  $D^2 f(\bar{x})(y, z) = 2\sigma(y * z)$ , this is a consequence of Lemma 3.12.  $\square$

Finally,

$$\left| A(Df(\bar{x} + y) - A^\dagger) z \right| \leq |A| (C^1(\bar{x})r + C^2 r^2),$$

and we are left to bound  $|A| C^1(\bar{x})$  and  $|A| C^2$ .

**Proposition 3.16.** *Let  $M$  be an integer satisfying (32) and (36). We define  $D^1 = (D_k^1)_{k \geq 0}$  component-wise by*

$$D_F^1(\bar{x}) \stackrel{\text{def}}{=} |A_m| C_F^1(\bar{x}) + \frac{2|\beta_{m-1}| \eta |\sigma| \alpha_m^s(m) \|\bar{x}\|_s}{C_1(1-\theta)\omega_m^{s+s_L}} |A_m|_{c_{m-1}}, \quad (46)$$

$$\begin{aligned} D_{m+k}^1(\bar{x}) \stackrel{\text{def}}{=} & \left( |A_m|_{r_{m-1}} C_F^1(\bar{x}) + \frac{2|\beta_{m-1}| |A_m|_{m-1, m-1} \eta |\sigma| \alpha_m^s(m) \|\bar{x}\|_s}{C_1(1-\theta)\omega_m^{s+s_L}} \right) \eta \frac{|\lambda_m|}{|\mu_m|} \theta^k \\ & + \frac{2\eta |\sigma| \alpha_m^s(m) \|\bar{x}\|_s}{C_1 \omega_{m+k}^{s+s_L}} \left( \sum_{l=0}^k \theta^{k-l} \left( \frac{m+k}{m+l} \right)^{s+s_L} + \frac{\theta}{1-\theta} \right), \quad \forall 0 \leq k < M, \end{aligned} \quad (47)$$

$$\begin{aligned} D_{m+M}^1(\bar{x}) \stackrel{\text{def}}{=} & \left( |A_m|_{r_{m-1}} C_F^1(\bar{x}) + \frac{2|\beta_{m-1}| |A_m|_{m-1, m-1} \eta |\sigma| \alpha_m^s(m) \|\bar{x}\|_s}{C_1(1-\theta)\omega_m^{s+s_L}} \right) \eta \frac{|\lambda_m|}{|\mu_m|} \theta^M \\ & + \frac{2\eta |\sigma| \alpha_m^s(m) \|\bar{x}\|_s}{C_1 \omega_{m+M}^{s+s_L}} \left( \chi + \frac{\theta}{1-\theta} \right), \end{aligned} \quad (48)$$

and

$$D_{m+k}^1(\bar{x}) \stackrel{\text{def}}{=} D_{m+M}^1(\bar{x}) \frac{\omega_{m+M}^s}{\omega_{m+k}^s}, \quad \forall k > M. \quad (49)$$

Then

$$|A| C^1(\bar{x}) \leq D^1(\bar{x}).$$

*Proof.* Thanks to (21),

$$\left( |A| C^1(\bar{x}) \right)_F \leq |A_m| C_F^1(\bar{x}) + |\beta_{m-1}| \left| U_I^{-1} L_I^{-1} C_I^1(\bar{x}) \right|_0 |A_m|_{c_{m-1}},$$

and using (33)

$$|U_I^{-1} L_I^{-1} C_I^1(\bar{x})|_0 \leq \frac{\eta \|C_I^1(\bar{x})\|_s}{C_1(1-\theta)\omega_m^{s+s_L}} \leq \frac{2\eta |\sigma| \alpha_m^s(m) \|\bar{x}\|_s}{C_1(1-\theta)\omega_m^{s+s_L}},$$

so that (46) holds. Still thanks to (21),

$$\begin{aligned} (|A| C^1(\bar{x}))_I &\leq |\lambda_m| |A_m|_{r_{m-1}} C_F^1(\bar{x}) |w_I| + |U_I^{-1} L_I^{-1} C_I^1(\bar{x})| + |\lambda_m| |\beta_{m-1}| |A_m|_{m-1, m-1} |U_I^{-1} L_I^{-1} C_I^1(\bar{x})|_0 |w_I| \\ &\leq |\lambda_m| \left( |A_m|_{r_{m-1}} C_F^1(\bar{x}) + \frac{2 |\beta_{m-1}| |A_m|_{m-1, m-1} \eta |\sigma| \alpha_m^s(m) \|\bar{x}\|_s}{C_1(1-\theta)\omega_m^{s+s_L}} \right) |w_I| + |U_I^{-1} L_I^{-1} C_I^1(\bar{x})|. \end{aligned}$$

Using (29) and (33), we get

$$\begin{aligned} (|A| C^1(\bar{x}))_{m+k} &\leq \left( |A_m|_{r_{m-1}} C_F^1(\bar{x}) + \frac{2 |\beta_{m-1}| |A_m|_{m-1, m-1} \eta |\sigma| \alpha_m^s(m) \|\bar{x}\|_s}{C_1(1-\theta)\omega_m^{s+s_L}} \right) \eta \frac{|\lambda_m|}{|\mu_m|} \theta^k \\ &\quad + \frac{2\eta |\sigma| \alpha_m^s(m) \|\bar{x}\|_s}{C_1 \omega_{m+k}^{s+s_L}} \left( \sum_{l=0}^k \theta^{k-l} \left( \frac{m+k}{m+l} \right)^{s+s_L} + \frac{\theta}{1-\theta} \right), \quad \forall 0 \leq k < M, \end{aligned}$$

so that (47) holds, and using (29) and (34), we get

$$\begin{aligned} (|A| C^1(\bar{x}))_{m+M} &\leq \left( |A_m|_{r_{m-1}} C_F^1(\bar{x}) + \frac{2 |\beta_{m-1}| |A_m|_{m-1, m-1} \eta |\sigma| \alpha_m^s(m) \|\bar{x}\|_s}{C_1(1-\theta)\omega_m^{s+s_L}} \right) \eta \frac{|\lambda_m|}{|\mu_m|} \theta^M \\ &\quad + \frac{2\eta |\sigma| \alpha_m^s(m) \|\bar{x}\|_s}{C_1 \omega_{m+M}^{s+s_L}} \left( \chi + \frac{\theta}{1-\theta} \right), \end{aligned}$$

so that (48) holds. As before, (49) follows from (36).  $\square$

We get similar results for the second order term.

**Proposition 3.17.** *Let  $M$  be an integer satisfying (32) and (36). Define  $D^2 = (D_k^2)_{k \geq 0}$  component-wise by*

$$D_F^2 \stackrel{\text{def}}{=} |A_m| C_F^2 + \frac{2 |\beta_{m-1}| \eta |\sigma| \alpha_m^s(m)}{C_1(1-\theta)\omega_m^{s+s_L}} |A_m|_{c_{m-1}},$$

$$\begin{aligned} D_{m+k}^2 &\stackrel{\text{def}}{=} \left( |A_m|_{r_{m-1}} C_F^2 + \frac{2 |\beta_{m-1}| |A_m|_{m-1, m-1} \eta |\sigma| \alpha_m^s(m)}{C_1(1-\theta)\omega_m^{s+s_L}} \right) \eta \frac{|\lambda_m|}{|\mu_m|} \theta^k \\ &\quad + \frac{2\eta |\sigma| \alpha_m^s(m)}{C_1 \omega_{m+k}^{s+s_L}} \left( \sum_{l=0}^k \theta^{k-l} \left( \frac{m+k}{m+l} \right)^{s+s_L} + \frac{\theta}{1-\theta} \right), \quad \forall 0 \leq k < M, \end{aligned}$$

$$D_{m+M}^2 \stackrel{\text{def}}{=} \left( |A_m|_{r_{m-1}} C_F^2 + \frac{2 |\beta_{m-1}| |A_m|_{m-1, m-1} \eta |\sigma| \alpha_m^s(m)}{C_1(1-\theta)\omega_m^{s+s_L}} \right) \eta \frac{|\lambda_m|}{|\mu_m|} \theta^M + \frac{2\eta |\sigma| \alpha_m^s(m)}{C_1 \omega_{m+M}^{s+s_L}} \left( \chi + \frac{\theta}{1-\theta} \right),$$

and

$$D_{m+k}^2 \stackrel{\text{def}}{=} D_{m+M}^2 \frac{\omega_{m+M}^s}{\omega_{m+k}^s}, \quad \forall k > M.$$

Then

$$|A| C^2 \leq D^2.$$

Finally we can sum up all the computations of this Sub-subsection and state the following result:

**Proposition 3.18.** *Let  $M$  be an integer satisfying (32) and (36). We define  $D^1$  (resp.  $D^2$ ) as in Proposition 3.16 (resp. Proposition 3.17) and let*

$$Z^2(r) \stackrel{\text{def}}{=} D^1(\bar{x})r + D^2r^2.$$

*Then for all  $y, z \in B_0(r)$*

$$A(Df(\bar{x} + y) - A^\dagger)z \leq Z^2(r).$$

Putting this together with Proposition 3.11, we end up with the following result:

**Proposition 3.19.** *Let  $M$  be an integer satisfying (32) and (36). Let*

$$Z(r) \stackrel{\text{def}}{=} Z^1(r) + Z^2(r).$$

*Then for all  $y, z \in B_0(r)$ ,*

$$|DT(\bar{x} + y)z| \leq Z(r).$$

### 3.4 The radii polynomials and interval arithmetics

All the work done up to now in Sections 2 and 3 can be summarized in the following statement:

**Theorem 3.20.** *Let  $s > 1$ , and  $s_L > 0$ . Assume that  $f$  is a map from  $\Omega^s$  to  $\Omega^{s-s_L}$  of the form  $f = \mathcal{L} + N$ , where  $\mathcal{L}$  is a tridiagonal operator satisfying (3), (4) and (5), and where the non linear part  $N$  is quadratic. Assume that for some  $m \geq 6$  we have computed an approximate zero of  $f$ , of the form  $\bar{x} = (\bar{x}_0, \dots, \bar{x}_{m-1}, 0, \dots, 0, \dots)$ , and  $D$  an approximate inverse of  $Df^{(m)}(\bar{x})$ . Consider*

$$T : \begin{cases} \Omega^s \rightarrow \Omega^s, \\ x \mapsto x - Af(x), \end{cases}$$

*where  $A$  is defined as in (21). Take  $M$  satisfying (32) and (36) and  $L \geq 0$  a computational parameter. Then the bound  $Y$  defined in Proposition 3.10 satisfies (26) and for all  $r > 0$ , the bound  $Z(r)$  defined in Proposition 3.19 satisfies (27).*

Now that we have found bounds  $Y$  and  $Z(r)$  that satisfy (26) and (27), we must find a radius  $r > 0$  such that  $\|Y + Z(r)\|_s < r$  in order to apply Theorem 3.1. By definition of the norm  $\|\cdot\|_s$ , it amounts to find an  $r > 0$  such that, for every  $k \geq 0$ , the radii polynomial  $P_k(r)$  satisfies

$$P_k(r) \stackrel{\text{def}}{=} Y_k + Z_k(r) - \frac{r}{\omega_k^s} < 0.$$

Note that since we constructed  $Y$  and  $Z$  in such a way that for every  $k \geq M$ ,

$$Y_{m+k} = Y_{m+M} \frac{\omega_{m+M}^s}{\omega_{m+k}^s} \quad \text{and} \quad Z_{m+k} = Z_{m+M} \frac{\omega_{m+M}^s}{\omega_{m+k}^s},$$

it is enough to find an  $r > 0$  such that for all  $0 \leq k \leq m + M$ ,  $P_k(r) < 0$ . In order to do so, we numerically compute, for each  $0 \leq k \leq m + M$ ,

$$I_k \stackrel{\text{def}}{=} \{r > 0 \mid P_k(r) < 0\},$$

and

$$I \stackrel{\text{def}}{=} \bigcap_{k=0}^{m+M} I_k.$$

If  $I$  is empty, then the proof fails, and we should try again with some larger parameters  $m$  and  $M$ . If  $I$  is non empty, we pick an  $r \in I$  and check rigorously, using the interval arithmetics package INTLAB [14], that for all  $0 \leq k \leq m+M$ ,  $P_k(r) < 0$ , which according to Theorem 3.1, proves that  $T$  defined in (25) is a contraction on  $B_s(\bar{x}, r)$ , thus yielding the existence of a unique solution of  $f(x) = 0$  in  $B_s(\bar{x}, r)$ .

## 4 An example of application

We present in this Section an example of equation, for which it is possible to apply the method developed in this paper. We first explain the link between the equation that we study (cf. (50) below) and the tridiagonal operator defined in Section 2. Then, we explain what are in this example the values of the various constants and parameters of our method.

Equations of the following form:

$$\begin{aligned} -(2 + \cos \xi)u''(\xi) + u(\xi) &= -\sigma u(\xi)^2 + g(\xi), \\ u'(0) = u'(\pi) &= 0, \end{aligned} \tag{50}$$

where  $g$  is a  $2\pi$ -periodic even smooth function, fall into the framework developed in Section 2. Consider indeed the cosine Fourier expansions of  $u$  and  $g$ :

$$u(\xi) = \sum_{k \in \mathbb{Z}} x_k \cos(k\xi), \quad g(\xi) = \sum_{k \in \mathbb{Z}} g_k \cos(k\xi).$$

Then, (50) can be rewritten as  $f(x) = 0$ , where

$$f_0(x) \stackrel{\text{def}}{=} x_0 + x_1 + \sigma (x * x)_0 - g_0,$$

and for all  $k \geq 1$ ,

$$f_k(x) \stackrel{\text{def}}{=} \frac{1}{2}(k-1)^2 x_{k-1} + (1+2k^2)x_k + \frac{1}{2}(k+1)^2 x_{k+1} + \sigma (x * x)_k - g_k. \tag{51}$$

We see that the linear part of (51) is, as in (3), given by

$$\mathcal{L}_k(x) = \lambda_k x_{k-1} + \mu_k x_k + \beta_k x_{k+1},$$

with

$$\mu_0 \stackrel{\text{def}}{=} 1, \quad \beta_0 \stackrel{\text{def}}{=} 1,$$

and for all  $k \geq 1$ ,

$$\lambda_k \stackrel{\text{def}}{=} \frac{1}{2}(k-1)^2, \quad \mu_k \stackrel{\text{def}}{=} (1+2k^2) \quad \text{and} \quad \beta_k \stackrel{\text{def}}{=} \frac{1}{2}(k+1)^2.$$

Let us fix some  $m \geq 2$ . With

$$C_1 = 2, \quad C_2 = 3 \quad \text{and} \quad \delta = \frac{1}{4} \frac{(m+1)^2}{m^2 + \frac{1}{2}},$$

we get

$$\forall k \geq 1, \quad \left| \frac{\lambda_k}{k^2} \right|, \left| \frac{\mu_k}{k^2} \right|, \left| \frac{\beta_k}{k^2} \right| \leq C_2,$$

together with

$$\forall k \geq m, \quad C_1 \leq \left| \frac{\mu_k}{k^2} \right| \quad \text{and} \quad \left| \frac{\lambda_k}{\mu_k} \right|, \left| \frac{\beta_k}{\mu_k} \right| \leq \delta.$$

We now focus on the example when

$$g(\xi) \stackrel{\text{def}}{=} \frac{1}{2} + 3 \cos(\xi) + \frac{1}{2} \cos(2\xi),$$

so that  $u(\xi) = \cos(\xi)$  is a trivial solution for  $\sigma = 0$ . We are going to use rigorous computations in order to prove the existence of solutions for  $\sigma \neq 0$ , and to compute these solutions.

Starting from  $\sigma = 0$ , we first use standard pseudo-arclength continuation techniques to numerically get some nontrivial approximate solutions for  $\sigma \neq 0$ . We computed 1250 different solutions (675 for  $\sigma > 0$  and 675 for  $\sigma < 0$ ). See Figure 2 for a diagram summing up those computations, where each point represents a solution of (50).

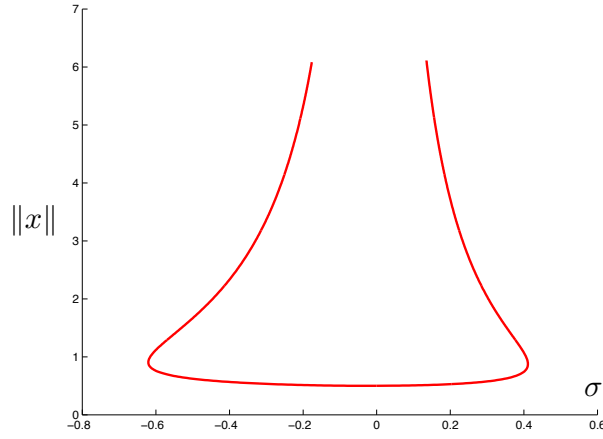


Figure 2: Branch of solutions of (50).

Then we use the rigorous computation method described in this paper to prove, for each numerical solution, the existence of a true solution in a small neighbourhood of the numerical approximation. We keep  $m = 20$  Fourier coefficients for the numerical computation, and use  $M = 20$  and the decay rate  $s = 2$  for the proof. The bounds of Lemma 3.12 as well as the error on  $\tilde{\omega}$  (35) are computed with  $L = 100$ . For each numerical solution, the proof is successful. The set  $I$  defined in Section 3.4 on which all radii polynomials should be negative always contains  $[4 \times 10^{-11}, 10^{-4}]$ , and we rigorously prove using interval arithmetics that they are indeed all negative for  $r = 10^{-10}$ . Hence the assumptions of Theorem 3.1 hold and as a consequence, within a ball of radius  $r = 10^{-10}$  in  $\Omega^s$  centered on the numerical approximation, there exists a unique solution to (50). Therefore the existence of the solutions represented in Figure 2 is rigorously proven, within a margin of error that is too small to be depicted. The codes used to perform the proofs can be found in [17].



Notice that existence of solutions of (50) could certainly have been obtained in different and more classical ways, for example using perturbative methods when  $\sigma$  is close to 0, or using a variational approach (that is, considering (50) as the Euler-Lagrange equation related to the critical points of a functional), or even using topological tools such as the Leray-Schauder theory. The advantage of our method is that it gives us more quantitative information than those approaches: indeed it enables to provide more than one solution for some values of  $\sigma$ , and, maybe more importantly, it gives a very precise localization of this (or these) solution(s) in terms of Fourier coefficients (something that looks very hard to obtain with qualitative PDEs methods).

## 5 Conclusion and Perspectives

A first interesting future direction of research would consist in adapting our approach to the rigorous computation connecting orbits of ODEs (using spectral methods). For instance, we would like to investigate the possibility of combining Hermite spectral methods with our approach to compute homoclinic orbits (e.g. see [15, 16]). Since the differential operator in frequency space of the Hermite functions is tridiagonal, adapting our method to this class of operator could lead to a new rigorous numerical method for connecting orbits.

It would also be interesting to adapt our method to the case of solutions belonging to the sequence space

$$\ell_\nu^1 = \{x = (x_k)_{k \geq 0} : \|x\|_\nu \stackrel{\text{def}}{=} \sum_{k \geq 0} |x_k| \nu^k < \infty\}$$

for some  $\nu \geq 1$ . With this choice of Banach space, we could use the fact that  $\ell_\nu^1$  is naturally a Banach algebra under discrete convolutions. This could greatly simplify the nonlinear analysis.

Note that assumption (5) requires the tridiagonal operator to have symmetric ratios between the diagonal terms and the upper and lower diagonal terms. This is a restriction that could hopefully be relaxed. Since many interesting problems involve tridiagonal operators with non symmetric ratios (as in the case of differentiation in frequency space of the Hermite functions), we believe that this is a promising route to follow.

Finally, generalizing our approach to problems with block-tridiagonal structures could also be a valuable project.

## Acknowledgement

The research leading to this paper was partially funded by the french “ANR blanche” project Kibord: ANR-13-BS01-0004.

## References

- [1] John P. Boyd. *Chebyshev and Fourier spectral methods*. Dover Publications Inc., Mineola, NY, second edition, 2001.
- [2] Allan Hungria, Jean-Philippe Lessard, and Jason D. Mireles-James. Radii polynomial approach for analytic solutions of differential equations: Theory, examples, and comparisons. To appear in *Math. Comp.*, 2015.

- [3] Piotr Zgliczyński and Konstantin Mischaikow. Rigorous numerics for partial differential equations: the Kuramoto-Sivashinsky equation. *Found. Comput. Math.*, 1(3):255–288, 2001.
- [4] Yasuaki Hiraoka and Toshiyuki Ogawa. Rigorous numerics for localized patterns to the quintic Swift-Hohenberg equation. *Japan J. Indust. Appl. Math.*, 22(1):57–75, 2005.
- [5] Marcio Gameiro and Jean-Philippe Lessard. Analytic estimates and rigorous continuation for equilibria of higher-dimensional PDEs. *J. Differential Equations*, 249(9):2237–2268, 2010.
- [6] Gábor Kiss and Jean-Philippe Lessard. Computational fixed-point theory for differential delay equations with multiple time lags. *J. Differential Equations*, 252(4):3093–3115, 2012.
- [7] S. Day, O. Junge, and K. Mischaikow. A rigorous numerical method for the global analysis of infinite-dimensional discrete dynamical systems. *SIAM J. Appl. Dyn. Syst.*, 3(2):117–160 (electronic), 2004.
- [8] Anthony W. Baker, Michael Dellnitz, and Oliver Junge. A topological method for rigorously computing periodic orbits using Fourier modes. *Discrete Contin. Dyn. Syst.*, 13(4):901–920, 2005.
- [9] Roberto Castelli and Jean-Philippe Lessard. Rigorous Numerics in Floquet Theory: Computing Stable and Unstable Bundles of Periodic Orbits. *SIAM J. Appl. Dyn. Syst.*, 12(1):204–245, 2013.
- [10] Maxime Breden, Jean-Philippe Lessard, and Matthieu Vanicat. Global Bifurcation Diagrams of Steady States of Systems of PDEs via Rigorous Numerics: a 3-Component Reaction-Diffusion System. *Acta Appl. Math.*, 128:113–152, 2013.
- [11] Philippe G. Ciarlet. *Introduction to numerical linear algebra and optimisation*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 1989. With the assistance of Bernadette Miara and Jean-Marie Thomas, Translated from the French by A. Buttigieg.
- [12] Donald E. Knuth. *The art of computer programming. Vol. 2*. Addison-Wesley Publishing Co., Reading, Mass., second edition, 1981. Seminumerical algorithms, Addison-Wesley Series in Computer Science and Information Processing.
- [13] Marcio Gameiro and Jean-Philippe Lessard. Efficient Rigorous Numerics for Higher-Dimensional PDEs via One-Dimensional Estimates. *SIAM J. Numer. Anal.*, 51(4):2063–2087, 2013.
- [14] S.M. Rump. INTLAB - INTerval LABoratory. In Tibor Csendes, editor, *Developments in Reliable Computing*, pages 77–104. Kluwer Academic Publishers, Dordrecht, 1999. <http://www.ti3.tu-harburg.de/rump/>.
- [15] Valeriy R. Korostyshevskiy and Thomas Wanner. A Hermite spectral method for the computation of homoclinic orbits and associated functionals. *J. Comput. Appl. Math.*, 206(2):986–1006, 2007.
- [16] Valeriy R. Korostyshevskiy. *A Hermite spectral approach to homoclinic solutions of ordinary differential equations*. ProQuest LLC, Ann Arbor, MI, 2005. Thesis (Ph.D.)—University of Maryland, Baltimore County.

- [17] M. Breden, L. Desvillettes and J.-P. Lessard. MATLAB codes to perform the proofs.  
<http://archimede.mat.ulaval.ca/jplessard/PseudoInverse>
- [18] J.-P. Lessard, J.D. Mireles James and J. Ransford. Automatic differentiation for Fourier series and the radii polynomial approach. *In preparation*.